DOCUMENT RESUME

ED 409 328 TM 026 644

AUTHOR Roberts, James S.; And Others
TITLE Comparative Validity of the Likert and Thurstone Approaches to Attitude Measurement.
PUB DATE Mar 97
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Attitude Measures; Attitudes; Comparative Analysis; *Error of Measurement; Item Response Theory; *Likert Scales; *Measurement Techniques; Research Methodology; *Validity
IDENTIFIERS *Thurstone Scales

ABSTRACT

Graded or binary disagree-agree responses to attitude statements are often collected for the purpose of attitude measurement. The empirical characteristics of these responses will generally be inconsistent with the analytical logic that forms the basis of the Likert attitude measurement technique (R. Likert, 1932). As a consequence, the Likert procedure can lead to invalid measurement of a select group of individuals. Likert attitude estimates can substantially misrepresent individuals with the most negative and most positive attitudes so that they appear to have more moderate opinions. In contrast, the Thurstone attitude measurement procedure (L. L. Thurstone, 1928) is generally more consistent with empirical characteristics of disagree-agree responses, and because of this superior consistency, Thurstone attitude scores do not suffer from this type of degraded validity. This paper highlights theoretical differences between the Likert and Thurstone approaches to attitude measurement and demonstrates how such differences can lead to discrepant attitude estimates for individuals with the most extreme opinions. Both simulated data and real data on attitude toward abortion are used to demonstrate this discrepancy. The results suggest that attitude researchers should, at the very least, devote more attention to the empirical response characteristics of items on a Likert attitude questionnaire. At most, these results suggest that other methods, like the Thurstone technique or one of its recently developed item response theory counterparts, should be used to derive attitude estimates from disagree-agree responses. (Contains 1 table, 12 figures, and 36 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
James Roberts
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 409 328

Comparative validity of the Likert and Thurstone approaches to attitude measurement[1].

James S. Roberts
Alcohol Research Center
Center for Drug and Alcohol Programs
Medical University of South Carolina

James E. Laughlin & Douglas H. Wedell
Department of Psychology
University of South Carolina

[1]This research was conducted while the first author was a postdoctoral fellow in the Division of Statistics and Psychometric Research at Educational Testing Service under the mentorship of Dr. John Donoghue. We would like to thank Dr. Donoghue for his helpful comments about this work. Correspondence should be send to James S. Roberts, Center for Drug and Alcohol Programs, Medical University of South Carolina, 171 Ashley Avenue, IOP - 4 North, Charleston, SC 29425-0742, e-mail: james_roberts@smtpgw.musc.edu.

TM026644

Abstract

Graded or binary disagree-agree responses to attitude statements are often collected for the purpose of attitude measurement. The empirical characteristics of these responses will generally be inconsistent with the analytical logic that forms the basis of the Likert attitude measurement technique. As a consequence, the Likert procedure can lead to invalid measurement of a select group of individuals - those individuals with the most extreme attitudes. Specifically, Likert attitude estimates can substantially misrepresent individuals with the most negative and most positive attitudes so that they appear to have more moderate opinions. In contrast, the Thurstone attitude measurement procedure is generally more consistent with empirical characteristics of disagree-agree responses, and because of this superior consistency, Thurstone attitude scores do not suffer from this type of degraded validity. This paper highlights the theoretical differences between the Likert and Thurstone approaches to attitude measurement, and demonstrates how such differences can lead to discrepant attitude estimates for individuals with the most extreme opinions. Both simulated data and real data on attitude toward abortion are used to demonstrate this discrepancy. The results suggest that attitude researchers should, at the very least, devote more attention to the empirical response characteristics of items on a Likert attitude questionnaire. At most, these results suggest that other methods, like the Thurstone technique or one of its recently developed item response theory counterparts, should be used to derive attitude estimates from disagree-agree responses.

Introductory texts often portray the Thurstone (1928) and the Likert (1932) approaches to attitude measurement as though both methods provide equally valid measures of attitude when individuals respond to a set of questionnaire items using a (binary or graded) disagree-agree response scale (Petty and Cacioppo, 1981; Mueller, 1986). This overly simplistic portrayal is fostered by studies which indicate that Likert and Thurstone attitude scores are typically correlated to a least a moderate degree ($.60 \leq r \leq .95$), regardless of whether responses to the same set of items are scored with the two procedures (Ferguson, 1941; Likert, 1932; Likert, Roslow & Murphy, 1934) or responses to independently constructed Likert and Thurstone questionnaires are compared (Edwards & Kenney, 1946; Flamer, 1983; Jaccard, Weber & Lundmark, 1975; Likert, 1932; Rhoads & Landy, 1973). Given these results, researchers have usually differentiated the two methods using other measurement criteria such as reliability and efficiency of scale construction. The general finding has been that Likert attitude scores exhibit either higher composite reliability (i.e., corrected split-half or corrected parallel forms reliability) or higher test-retest reliability as compared to Thurstone attitude scores (Seiler & Hough, 1970). Additionally, the general perception is that the Likert technique is easier and more efficient to carry out than the Thurstone technique, primarily because the former method does not require a judgment group to produce item scale values (Barclay & Weaver, 1962; Edwards & Kenny, 1946; Mueller, 1986). These two features may account for the relatively superior popularity of the Likert procedure for attitude measurement (Petty & Cacioppo, 1981).

Although previous studies have suggested that Likert and Thurstone attitude scores will be linearly related to at least a moderate extent, they do not convincingly demonstrate that the two scores both measure the latent attitude with the same degree of validity. The relationship between Likert scores and true attitudes could still differ systematically from the corresponding relationship found for Thurstone scores whenever the correlation between the two types of scores is only moderately high. Therefore, distinctions between the two procedures might still be made with regard to their respective validities.

We argue against the idea that the Thurstone and Likert methods generally yield comparably valid estimates of true attitudes, and for the idea that the methods should not be treated as equally applicable in traditional attitude measurement situations. Instead, the appropriate application of either method depends on the item response process that subjects use when endorsing attitude items. We also argue that in those traditional situations where subjects respond to attitude items using a graded or binary disagree-agree response scale, the empirical response process generally favors the use of the Thurstone procedure as opposed to the Likert procedure. Moreover, we use both simulated and real data to illustrate how the application of the Likert procedure in these situations can yield invalid measures for individuals with the most extreme attitudes. In contrast, the validity of the Thurstone procedure does not degenerate in these situations.

Review of the Thurstone and Likert Approaches

The Thurstone Approach

The classic Thurstone approach to attitude scale construction involves two main stages. In

the first stage, a large number of attitude statements are written to span the entire range of possible opinions, and these items are scaled with regard to their unfavorability or favorability towards a given attitude object. There are several Thurstonian techniques for scaling attitude items including pairwise comparisons (Thurstone, 1927abc), equal appearing intervals (Thurstone & Chave, 1929), and successive intervals (Safir, 1937). All of these methods require a group of subjects to make favorability judgments about each item (or each pair of items), and all three methods yield a set of item scale values which indicate how favorably or unfavorably the item's sentiment reflects the attitude object. Those items with scale values having large standard errors are discarded from the pool of items under consideration. In the second stage, subjects are asked to indicate which attitude statements they agree with. Attitude estimates are developed for each individual by computing the mean (or median) scale value associated with endorsed item, and then these attitude estimates are used to develop empirical operating characteristic curves for each item. The final Thurstone scale is limited to "relevant" items with scale values that are more or less uniformly distributed across the attitude continuum. A relevant item is one that attracts endorsements primarily from subjects whose attitudes are comparable to the sentiment expressed by the item.

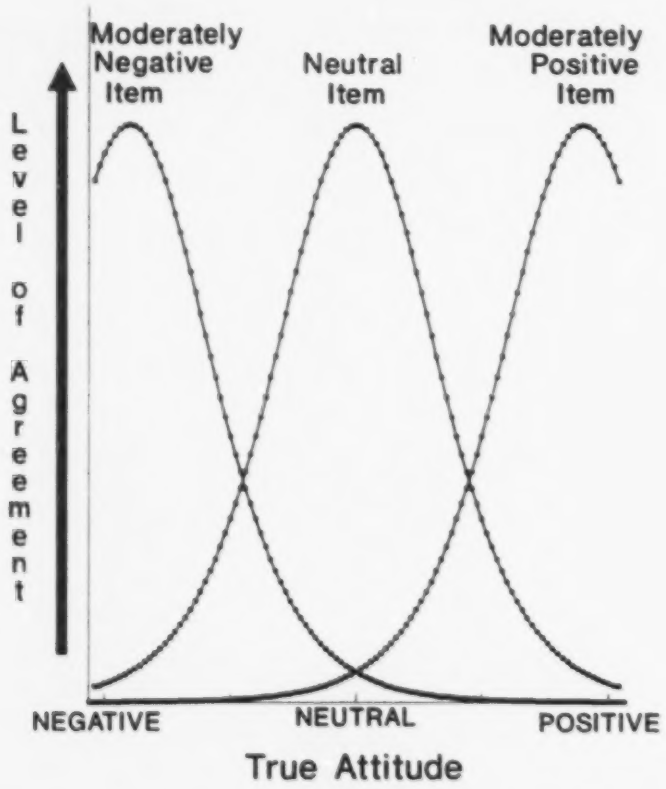

Figure 1. Theoretical item characteristic curves associated with an unfolding model. From (upper) left to right, the curves correspond to a moderately negative item, a neutral item and a moderately positive item

The operating characteristic of a relevant Thurstone item reflects Coombs' (1964) notion of an ideal point process - a process where the individual endorses an attitude item to the extent that it reflects the individual's own opinion. Responses resulting from an ideal point process are best analyzed with some form of unfolding model in which the probability of endorsement is a function of the proximity between an individual and an item on the underlying attitude continuum. Moreover, by limiting the final scale to only relevant items, the Thurstone procedure can be regarded as a type of unfolding model (Andrich, 1988, 1996; Andrich & Luo, 1993; Roberts, 1995). Figure 1 illustrates the theoretical item characteristic curves (ICCs) predicted from an unfolding model. For example, a neutral item should be endorsed most by individuals with relatively moderate attitudes, and it should be endorsed less frequently by persons with more extreme attitudes in either direction. In contrast, a moderately positive item should be endorsed most by individuals with moderately positive attitudes, and it should be endorsed less by those with neutral opinions, and even less so by persons with negative attitudes. Additionally, because the unfolding model operates on the basis of the absolute distance between an individual and an item on the continuum, those persons with extremely positive attitudes may exhibit relatively less agreement with a moderately positive item because it fails to reflect the extremity of their opinions. A moderately negative item would be characterized by the opposite pattern of responding as shown in Figure 1.

The Likert Procedure

The Likert procedure attempts to measure individual attitudes without deriving the locations of attitude items on the underlying continuum (i.e., without deriving item scale values). When constructing a Likert scale, a large number of preliminary items are developed such that each item expresses a clearly negative or positive opinion - neutral items are avoided. Subjects are generally asked to indicate how much they disagree or agree with each item using a graded disagree-agree scale. Responses to negatively worded items are reverse scored, and then all responses are subjected to a variety of analyses which attempt to identify the most discriminating, homogeneous and reliable items. These techniques may involve calculating discrimination indices, item-total correlations, item-deleted alpha coefficients, and/or principal components. The final scale is limited to a reasonably small set (generally 20 or fewer) items that appear optimal with regard to one or more of these criteria.

Likert never provided a theoretical model to justify his method[2], but nonetheless, the procedures commonly used to select final scale items are consistent with the idea of a dominance response process (Coombs, 1964). In a dominance response process, an individual endorses an item to the extent that the individual is located above the item on the underlying continuum. Responses from a dominance process are generally analyzed with some form of cumulative model in which the probability of endorsement increases as the signed distance between the individual and the item on the attitude continuum increases. Several researchers have noted that the Likert procedure is, in a functional sense, a type of cumulative model (Andrich, 1996; Green, 1954; Roberts, 1995). Figure 2 illustrates the theoretical ICCs associated with a general cumulative

---

[2]The use of classical test theory to justify the procedure occurred years after Likert's original proposal.

model. Specifically, individuals are expected to agree with a positively worded item to the extent that their attitudes are more positive than (i.e., dominate) the sentiment expressed by the item. Conversely, individuals are expected to endorse a negatively worded item to the extent that their attitudes are more negative than the opinion expressed by the item. Recall that in the Likert procedure, responses to negatively worded items are reverse scored. As a result of this rescoring, the monotonically decreasing ICC shown in Figure 2 is reflected along the vertical axis, and yields a monotonically increasing characteristic curve.

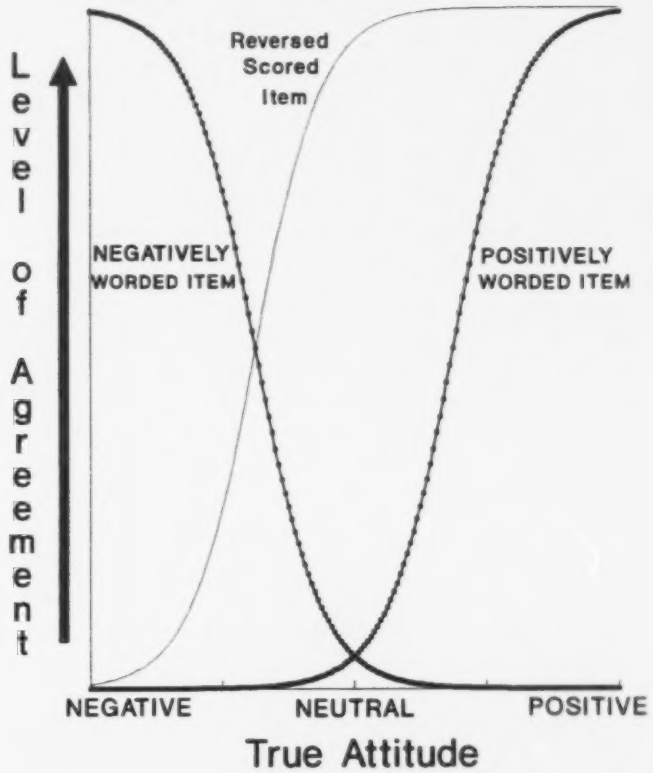

Figure 2. Theoretical item characteristic curves associated with a cumulative model. From (upper) left to right, the curves represent a negatively worded item, the same negatively worded item after reverse scoring item responses, and a positively worded item.

The Empirical Response Process

Several researchers (Andrich, 1996; Roberts, 1995; van Schuur & Kiers, 1994) have argued that subjects generally use some type of ideal point response process when they respond to attitude statements using either a graded or binary disagree-agree response scale. This perspective results from the fact that empirical ICCs derived from such statements typically resemble those in Figure 3. The 10 items shown in Figure 3 were designed to measure attitudes

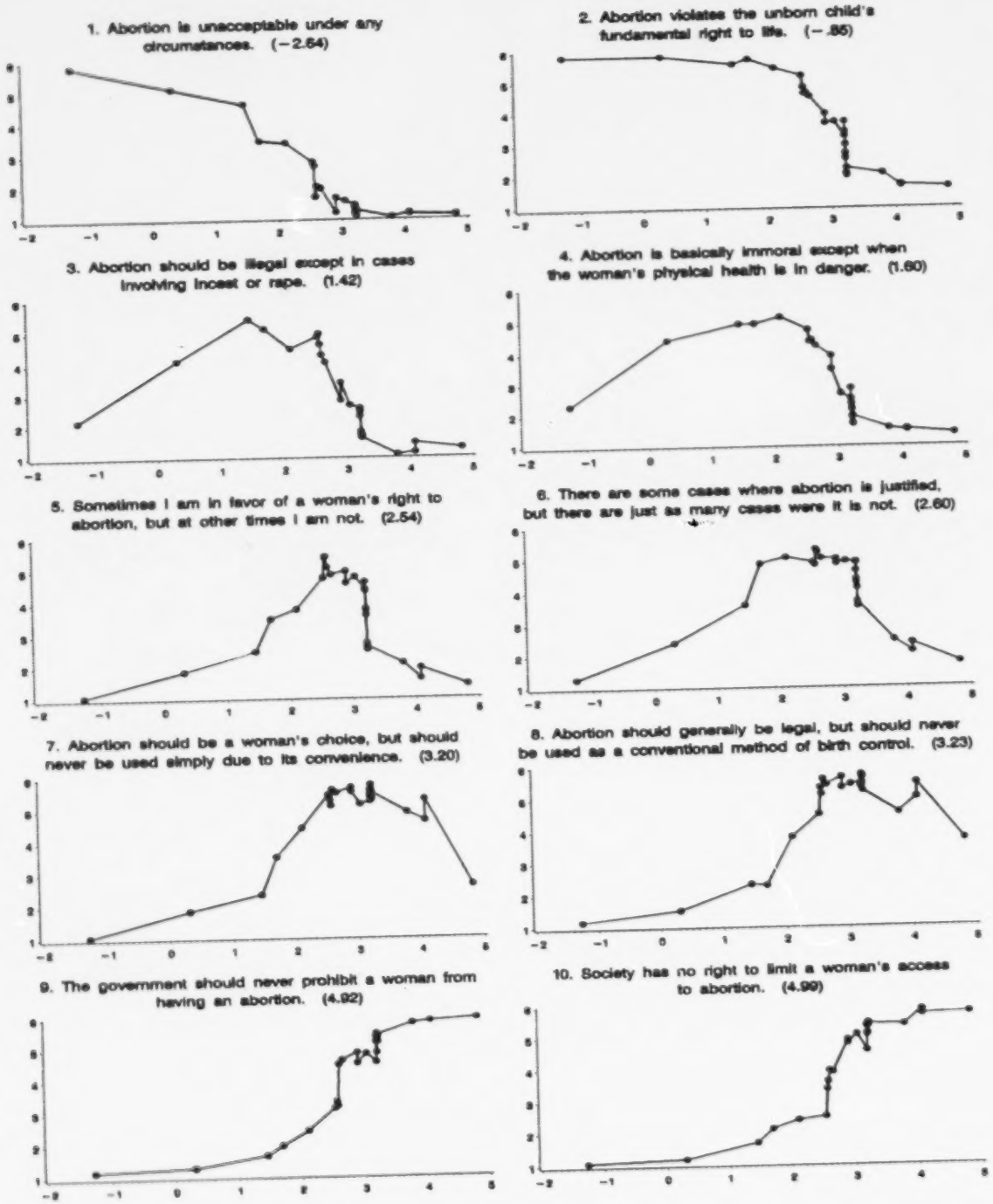
1. Abortion is unacceptable under any circumstances. (−2.64)
2. Abortion violates the unborn child's fundamental right to life. (−.85)
3. Abortion should be illegal except in cases involving incest or rape. (1.42)
4. Abortion is basically immoral except when the woman's physical health is in danger. (1.60)
5. Sometimes I am in favor of a woman's right to abortion, but at other times I am not. (2.54)
6. There are some cases where abortion is justified, but there are just as many cases were it is not. (2.60)
7. Abortion should be a woman's choice, but should never be used simply due to its convenience. (3.20)
8. Abortion should generally be legal, but should never be used as a conventional method of birth control. (3.23)
9. The government should never prohibit a woman from having an abortion. (4.92)
10. Society has no right to limit a woman's access to abortion. (4.99)

Figure 3. Empirical item characteristic curves associated with 10 items designed to measure attitudes toward abortion. Each vertical axis denotes the mean level of observed agreement (1="Strongly Disagree",... , 6="Strongly Agree"), whereas each horizontal axis denotes the mean Thurstone attitude score. Means were calculated within homogeneous Thurstone attitude score groups composed of approximately 30 respondents per group.

BEST COPY AVAILABLE

toward abortion. The corresponding ICCs were generated by first scaling the attitude statements using the successive intervals procedure. Scale values derived from this procedure were based on the statement favorability judgments of 303 subjects. Additionally, graded-disagree agree responses from 781 subjects were obtained for each item using a 6-point response scale where 1=Strongly Disagree, 2=Disagree, 3=Slightly Disagree, 4=Slightly Agree, 5=Agree, and 6=Strongly Agree. Each subject's attitude score was estimated from the median scale value associated with those items the subject agreed with to at least some extent (i.e., estimates were derived using the Thurstone procedure). Subjects were then sorted into one of 26 homogeneous attitude score groups with approximately 30 individuals per group. The mean attitude score for each group is portrayed on the horizontal axis and the average item response for each group is given on the vertical axis. The ICCs are arranged on the basis of item content (from very negative, to moderate, to very positive sentiments), and the corresponding Thurstone scale values are given parenthetically beside each item.

The most important feature about the ICCs in Figure 3 is that they are more consistent with an ideal point response process than a dominance response process. Specifically, the ICCs for the two extremely negative statements are more or less a monotonically decreasing function of estimated attitude such that individuals with the most negative opinions endorse these statements the most. However, the ICCs begin to exhibit a marked degree of nonmonotonic behavior as the corresponding attitude statements become more moderate in content. We refer to this nonmonotonic behavior as "folding". The "folding" of ICCs is first apparent with the moderately negative statements where those individuals with moderately negative attitudes show the highest levels of endorsement, but those subjects to either side of this attitudinal position exhibit less and less agreement. The folding becomes marked when considering neutral items. In those cases individuals with relatively neutral attitudes exhibit the most agreement, and those subjects with attitudes that are more extreme in either a negative or positive direction reveal substantially lower levels of agreement. The ICCs for the moderately positive items are opposite in appearance from those for moderately negative items, albeit somewhat less folded in this case. Furthermore, the ICCs for the extremely positive items are more or less monotonically increasing with estimated attitude. When taken together, these ICCs suggest that some type of ideal point response process is operating when subjects respond to items like those in Figure 3.

Theoretical Differences Between Methods

If disagree-agree responses to attitude statements generally follow from some type of ideal point process, then why has the Likert procedure performed reasonably well in these situations? This question can best be answered by comparing the ICCs for a moderately positive item under both an unfolding model and a cumulative model. This comparison is shown in Figure 4. The degree of correspondence between theoretical ICCs under both models is considerable for all but the most extremely positive regions of the attitude continuum. For less extreme attitude positions, both models make similar predictions about how individuals will respond to the item. However, the two models make divergent predictions for individuals with the most positive attitudes. The unfolding model suggests that individuals with extremely positive attitudes will

begin to agree less with a moderately positive item because the item does not reflect the extremity of their opinion well enough. In contrast, the cumulative model suggests that individuals with extremely positive attitudes will endorse a moderately positive item as much or more so than individuals with less positive attitudes. Consequently, the unfolding model predicts that individuals with extreme attitudes will exhibit less agreement than that predicted by the cumulative model. A similar scenario can be constructed to describe how individuals with the most negative attitudes would respond to moderately negative items under both models.

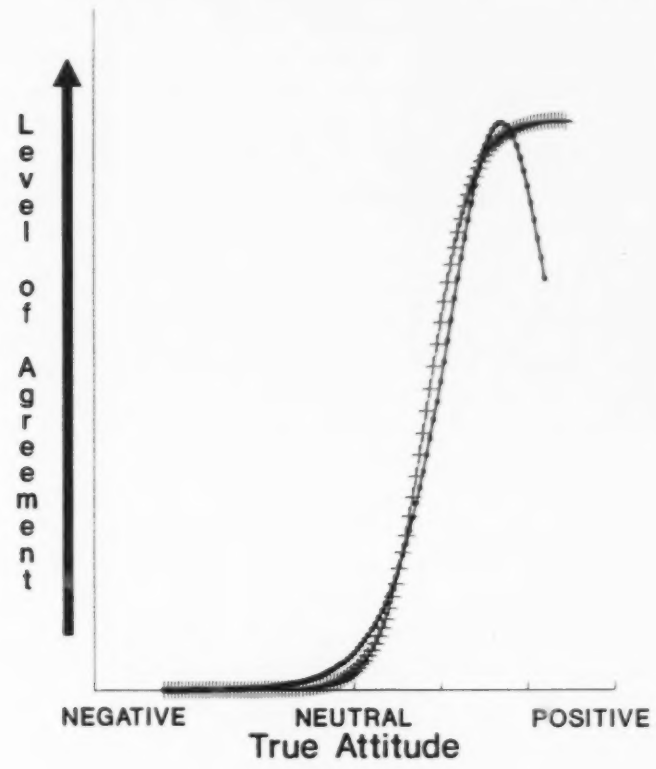

Figure 4. Comparison of the theoretical item characteristic curves associated with an unfolding model (denoted by squares) and a cumulative model (denoted by plus signs).

Given the premise of an ideal point response process, the Likert procedure performs at least reasonably well because neutral items that exhibit the most nonmonotonic ICCs are not typically included on the scale (Andrich, 1996; Edwards & Kenney, 1946; Ferguson, 1941, Roberts, 1995). Instead, the scale is generally limited to moderately extreme and extreme items that exhibit relatively small amounts of nonmonotonic behavior, if any. The item selection procedures that are used to develop Likert scales help alleviate the most offending neutral items from consideration. Moreover, instructions for developing suitable items for a Likert scale may explicitly suggest that

neutral items be avoided (Mueller, 1986). After these traditional scale construction techniques are applied, the selected scale items will generally exhibit a high degree of monotonicity, yet there will often be some nonmonotonic behavior in the extreme attitude regions as illustrated in Figure 4. As we shall see, this nonmonotonic behavior can lead to problems when obtaining measurements with the Likert technique.

Implications of the Theoretical Differences

The theoretical item response characteristics associated with the unfolding and cumulative models can be quite similar when considering the behavior of moderately extreme items in non-extreme regions of the attitude continuum. However, these models can produce substantially different expectations in more extreme regions of the attitude continuum. These theoretical differences lead to some specific predictions with regard to comparisons of Likert and Thurstone attitude scores derived from disagree-agree responses:

1) Likert and Thurstone scores should be monotonically related to each other in those cases where individual attitudes are not too extreme relative to the items under study.

2) Likert and Thurstone scores should be nonmonotonically related in situations where individual attitudes are substantially more extreme than the items in question.

3) When this nonmonotonicity occurs, Likert scores will incorrectly suggest that individuals with the most extreme Thurstone scores actually have more moderate opinions.

In the pages that follow, we use both simulated and real data to provide support for these hypotheses.

A Simulated Example

Method

The responses of 200 simulees were generated to a series of 59 attitude items using the Graded Unfolding Model (Roberts, 1995; Roberts & Laughlin, 1996ab). This model assumes an ideal point response process and produces ICCs similar to those shown in Figure 1. Responses were on a 6-point scale where 1 represented the strongest level of disagreement and 6 represented the strongest level of agreement. Each simulee's "true attitude" was randomly sampled from an N(0,2) distribution. Scale values for the 59 items ranged from -4.35 to +4.35 and divided the true attitude continuum into equally spaced segments of size .15. These person and item parameters were similar to those used in past evaluations of unfolding models. Each simulee's responses to the 59 items were independently replicated 100 times using the same nominal parameters.

On the first replication, responses to the 59 items were subjected to a principal components

analysis, and the items with the largest absolute pattern loadings were chosen to form an optimal 20-item Likert scale under the constraint that 10 items were from the negative side of the true attitude continuum and 10 items were from the positive side. The negative items were reversed scored, and the Likert attitude estimate for a given simulee was simply the sum of the 20 scored responses. These same 20 items were used to generate Likert scores for each simulee on subsequent replications.

A 20-item Thurstone scale was constructed by choosing items that spanned the latent attitude continuum from -4.05 to +4.05 in equal intervals of .45. (One of the 20 items was arbitrarily located at .15). Thurstone attitude scores were computed by averaging the true scale values associated with items endorsed by a given simulee. Endorsement was defined as a response of 4 or more.

At the end of each replication, a given simulee had a Likert score and Thurstone score based on the 20 items that were deemed optimal for the method in question. Each simulee's Likert scores were averaged across the 100 replications and the same was done for the corresponding Thurstone scores. Average Likert and Thurstone scores were then compared to true attitudes.

Results

Selection of Likert Items

Figure 5 illustrates the pattern loadings on the first principal component for the 59 items from the first replication. The 10 negative items identified as optimal by the principal components procedure had true scale values within the interval of [-3.00, -1.65]. Similarly, the optimal positive items were located in the [+1.95, +3.30] interval on the attitude continuum. Thus, the selected items were from moderately extreme regions of the attitude continuum as opposed to the most extreme regions. Consequently, the ICCs associated with the selected items exhibited more folding than those associated with the most extreme, unselected items. This fact is illustrated in Figure 6 which contrasts the empirical ICC for the most positive item selected for the Likert scale with that for the most positive item in the initial item pool. The most extreme item from the pool exhibited an essentially monotonic ICC. In contrast, the ICC associated with most positive item selected for the Likert scale exhibited a substantial degree of folding in the extremely positive regions of the attitude continuum. Even with these clearly nonmonotonic items, the Likert scale yielded a Cronbach's alpha of .96, and all corrected item-total correlations were greater than .70.

Relationships Among True Attitudes, Likert Estimates and Thurstone Estimates

Figure 7 illustrates the relationships found between mean Likert scores, mean Thurstone scores, and true attitudes. The relationship between average Thurstone scores and true attitudes was monotonically increasing throughout the simulated continuum. However, the relationship between mean Likert scores and true attitudes was markedly nonmonotonic. As expected, the nonmonotonicity was confined to those simulees with extreme true attitudes. In those instances, Likert scores consistently suggested more moderate attitude positions when, in fact, the corresponding true attitudes were the most extreme.

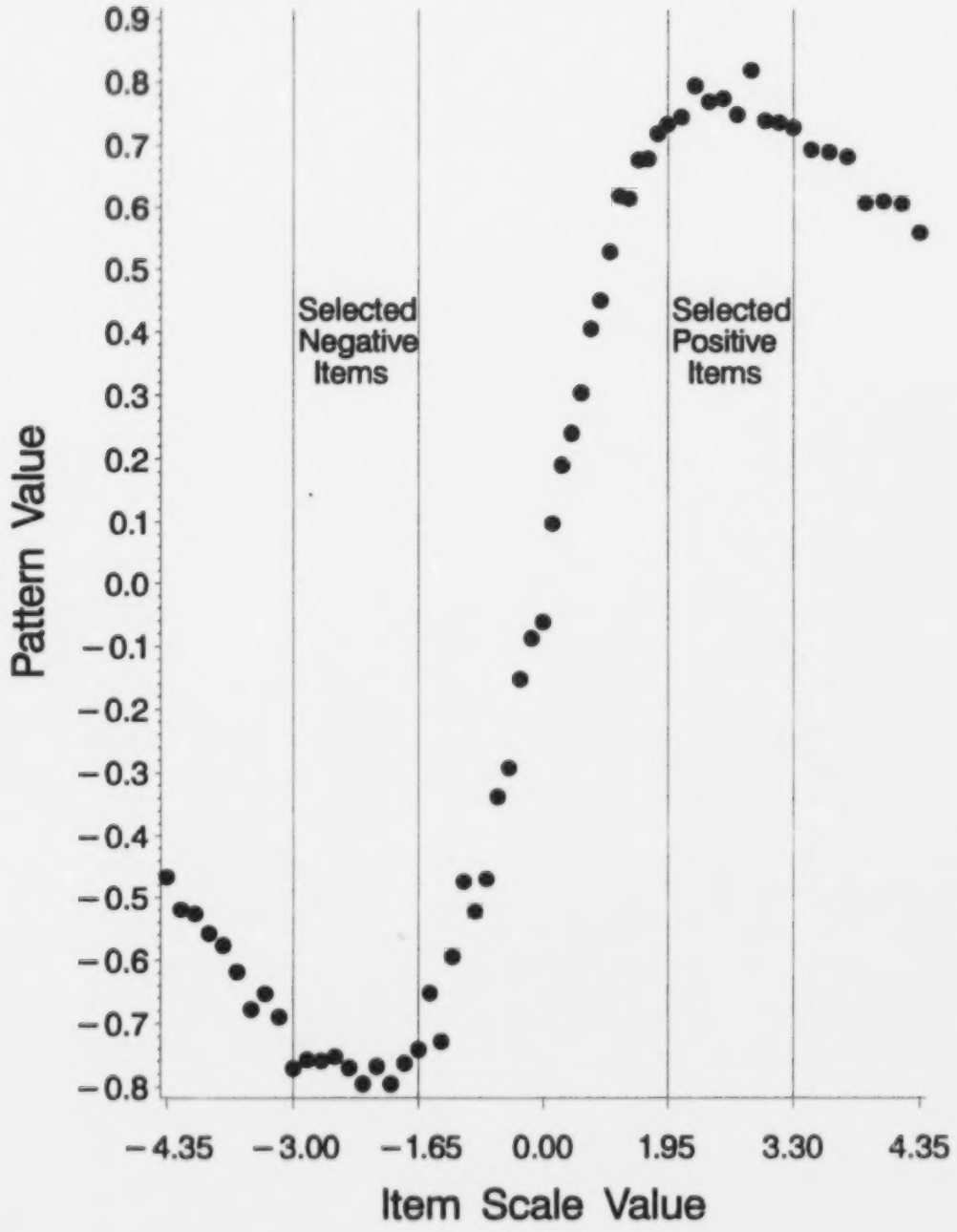

Figure 5. Pattern loadings from the first principal component of the interitem correlation matrix associated with simulated responses to 59 items. Loadings are plotted against the true scale value for each item.

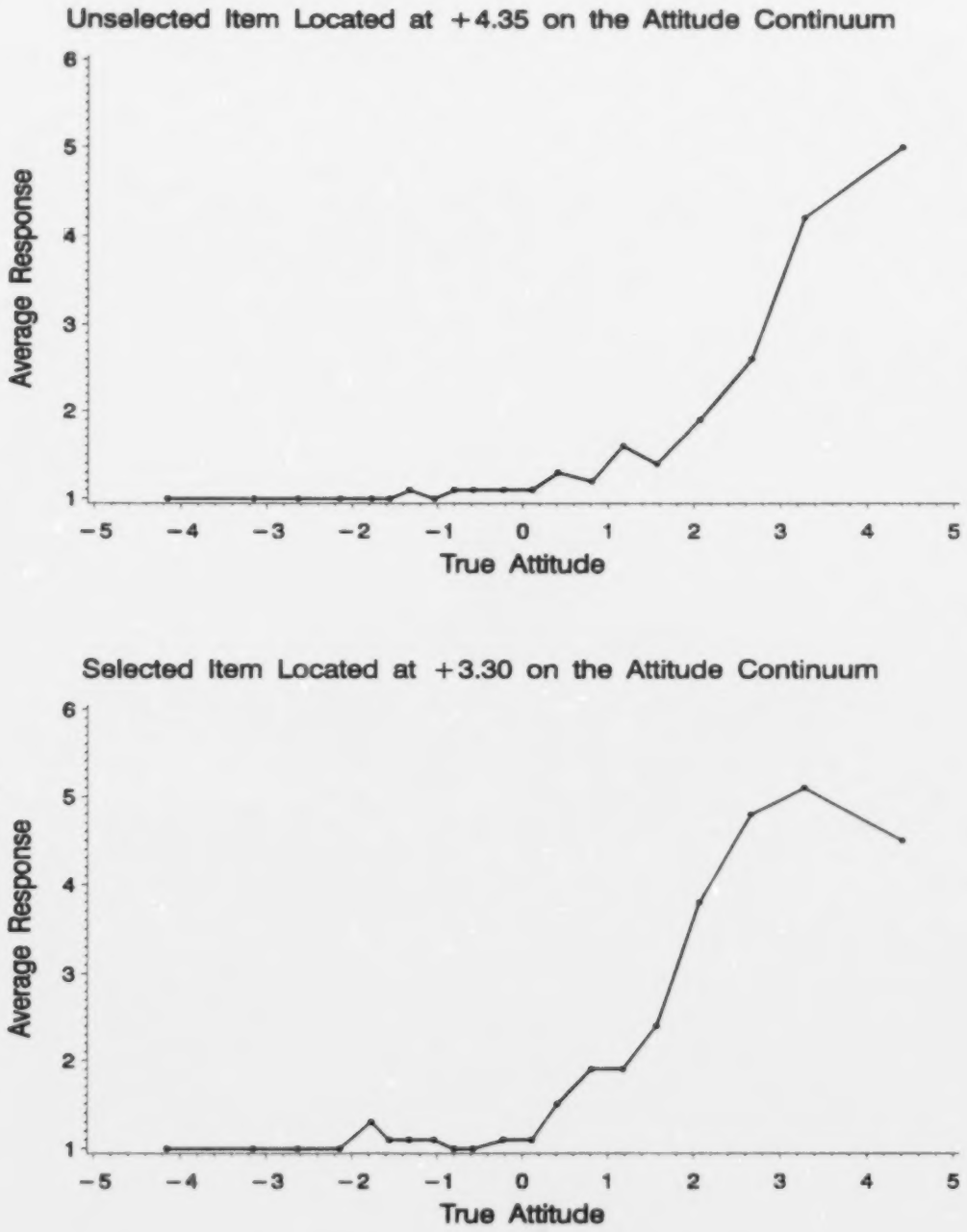
Unselected Item Located at +4.35 on the Attitude Continuum
Average Response
True Attitude
Selected Item Located at +3.30 on the Attitude Continuum
Average Response
True Attitude

Figure 6. Empirical item characteristic curves associated with the most extreme (unselected) item from the initial item pool and the most extreme (selected) item from the optimal Likert scale. Each point on a given curve represents the mean response and mean true attitude of 10 simulees.

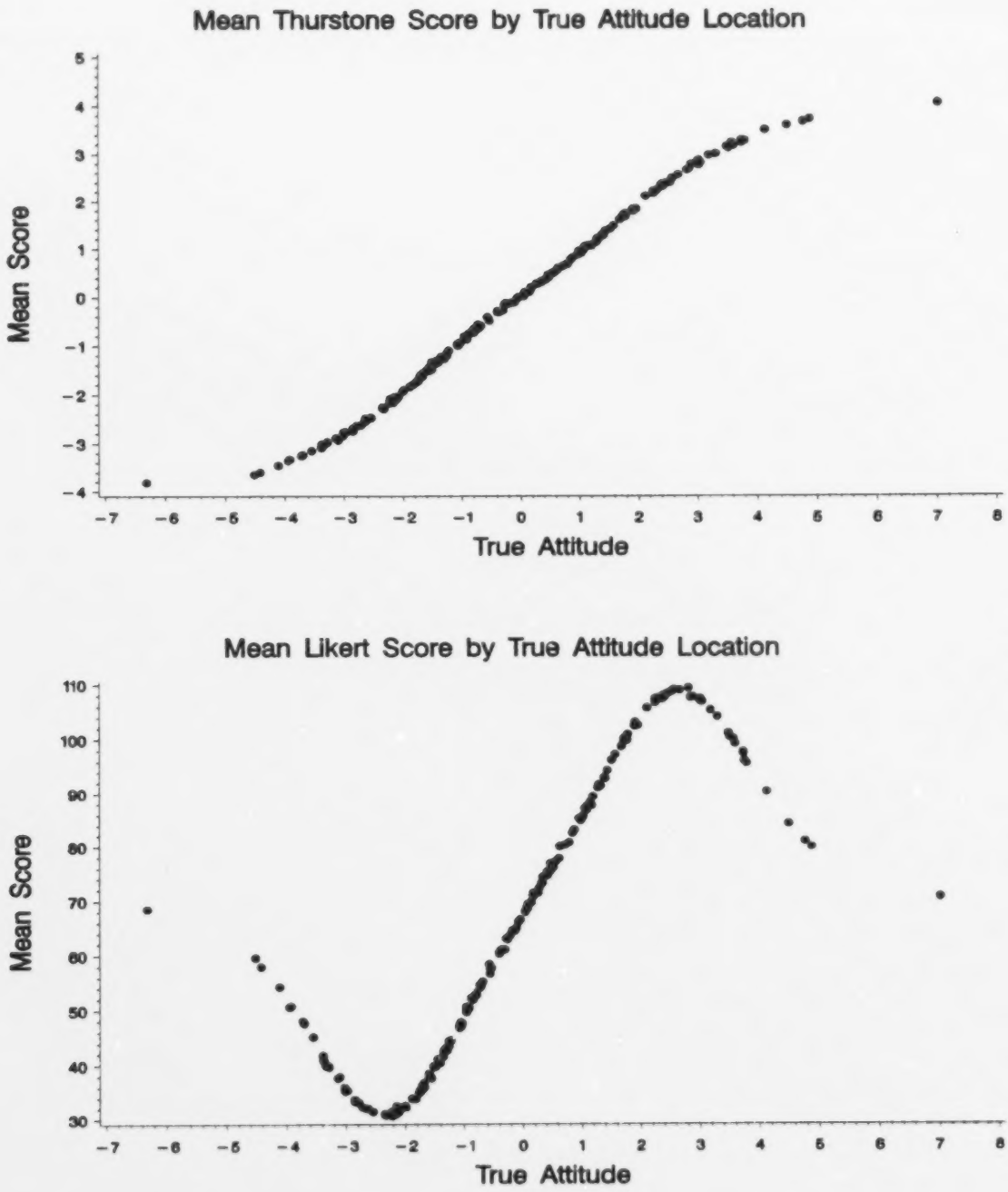

Figure 7. The relationship between true attitude and mean Thurstone attitude score (top panel), and true attitude and mean Likert score (bottom panel). Means were calculated across 100 replications.

Figure 8 directly compares the mean Likert and Thurstone estimates. There was a nonmonotonic relationship between the two sets of attitude estimates such that those simulees with the most extreme Thurstone scores had corresponding Likert scores that were indicative of more moderate attitude positions. This gave rise to an elongated S-shaped function relating the two measures.

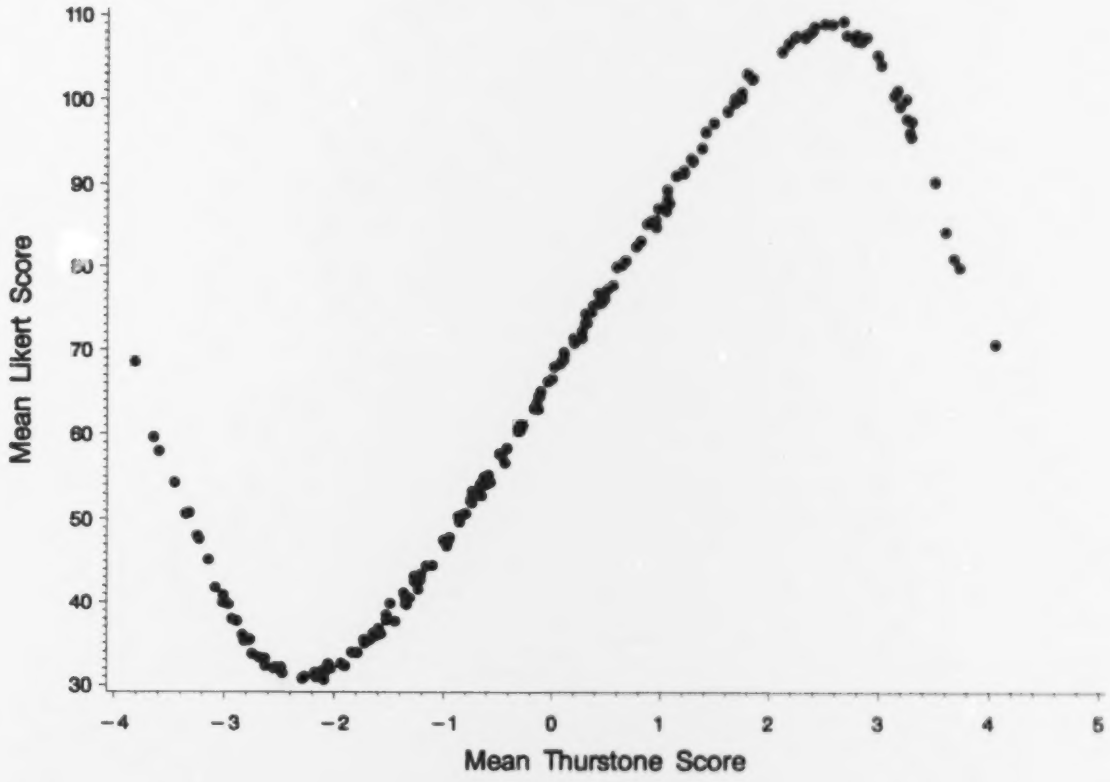

Figure 8. The relationship between mean Likert and mean Thurstone attitude scores computed across 100 replications.

An Example with Real Data

Method

Graded disagree-agree responses to the 10 items in Figure 3 were obtained from 781 subjects. Of theses subjects, 750 were undergraduate students at the University of South Carolina and 31 were members of special interest groups who were known in advance to be for or against the legal status of abortion. Responses to each item were on a 6-point scale where 1=Strongly Disagree, 2=Disagree, 3=Slightly Disagree, 4=Slightly Agree, 5=Agree, and 6=Strongly Agree.

A series of item analyses were conducted to determine which of the 10 items would be suitable for a Likert scale. The results suggested that the 8 non-neutral items corresponding to the two upper and two lower panels of Figure 3 would suffice. Together, these 8 items produced corrected item-total correlations that ranged from .45 to .75 and the corresponding pattern loadings from the first principal component ranged between .58 and .83. Cronbach's alpha for the 8-item scale was equal to .87. Thus, these item analysis indices suggested that the resulting 8-item Likert scale was acceptable for applied attitude research.

The ICCs in Figure 3 suggested that all the items would be suitable for a Thurstone attitude scale. However, the resulting Thurstone scale was not optimal in the sense that the items were not equally spaced across the attitude continuum, although they did adequately represent alternative regions of that continuum. To compensate for the unequal item spacing, each individual's attitude score was derived by computing the median (as opposed to the mean) of scale values associated with items that the individual endorsed to at least some extent.

Results

Figure 9 illustrates the relationship between each individual's Likert and Thurstone scores. The data were smoothed using a cubic spline method (Reinsch, 1967) and the resulting curve is also shown in the plot. The correlation between Likert and Thurstone scores was .80, and thus, the degree of linear association between the two measures was within the range reported in previous comparison studies (i.e., $.60 \leq r \leq .95$). The data were, nonetheless, consistent with the pattern expected under an ideal point hypothesis. Namely, the Likert and Thurstone scores were monotonically (if not linearly) related for those individuals with non-extreme Thurstone attitude estimates. However, the relationship between the two sets of estimates became nonmonotonic in the extreme regions of the Thurstone continuum. Specifically, the Likert method suggested that individuals with the most extreme Thurstone attitude scores actually had more moderate opinions relative to the other individuals in the sample.

In addition to smoothing the data with a spline function, the mean values of attitude estimates were also calculated in an effort to corroborate the primary trends in the data. Specifically, the data were ranked on the basis of Thurstone estimates and then partitioned into 26 relatively homogenous attitude groups with approximately 30 individuals per group. The means of the Likert and Thurstone estimates were computed within each of these groups. The resulting means are shown in Figure 10. Again, the pattern predicted from the ideal point hypothesis is evident, although weaker. The relationship between the two estimates was nonmonotonic in the extreme portions of the Thurstone attitude continuum. Moreover, the Likert method suggested that those individuals with the most extreme attitudes actually had more moderate opinions relative to the other subjects in the sample.

The nature of the nonmonotonic relationship between Likert and Thurstone scores became even more apparent when looking separately at the negatively worded and positively worded items from the scale. The 4 negatively worded items had a Cronbach's alpha value of .82 and the

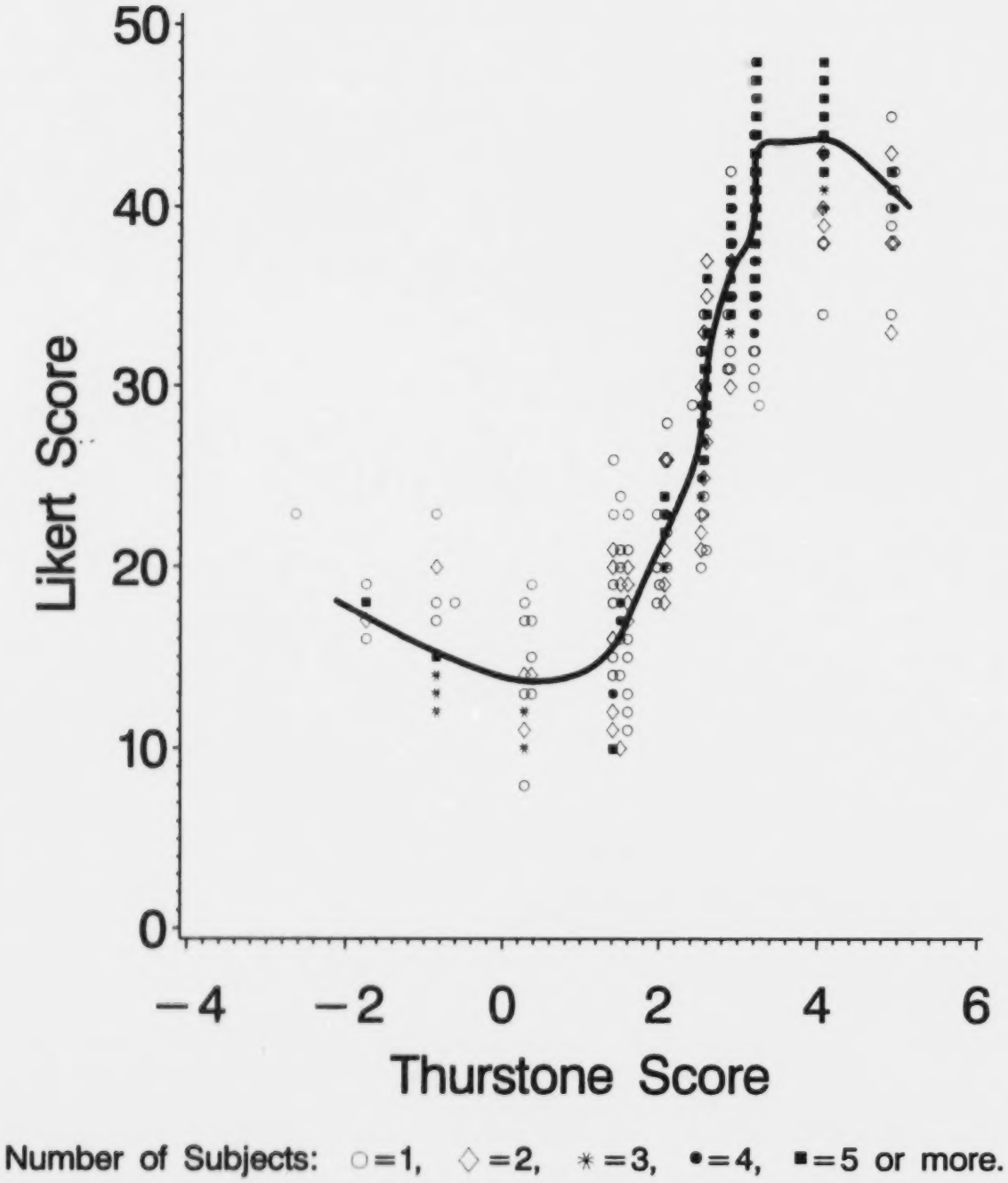

Figure 9. The relationship between observed Likert and Thurstone scores derived from 781 respondents. The number of individuals at a given point on the graph is indexed by the symbol associated with that point. The solid curve represents the value predicted by a cubic spline smoothing function applied to the data.

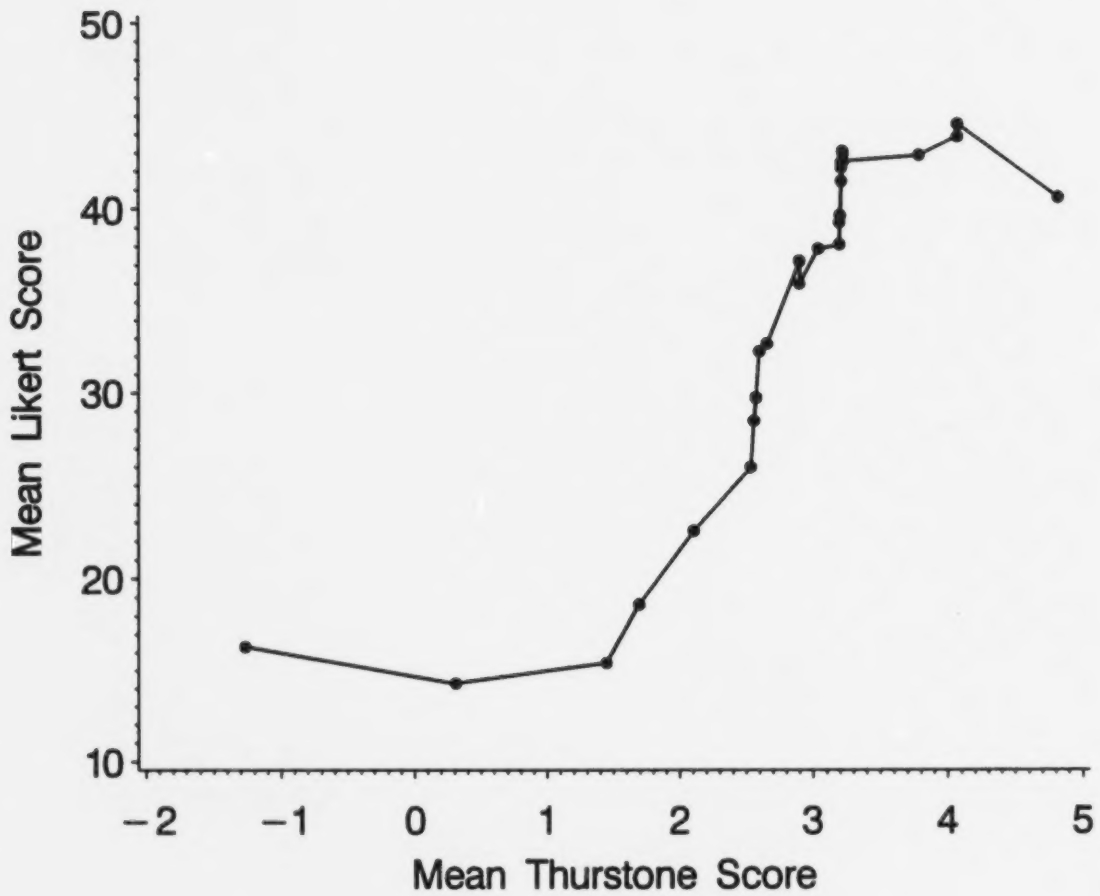

Figure 10. The relationship between mean Likert and mean Thurstone attitude scores. Means were calculated within relatively homogeneous Thurstone attitude score groups. Each mean was based on approximately 30 responses.

corrected item-total correlations ranged from .53 to .72. Figure 11 illustrates the relationship between the means of the original Thurstone attitude estimates and the Likert scores derived from the 4 negatively worded items. (Again, these means were calculated within each of the 26 attitude groups described previously.) The relationship among mean attitude scores is monotonic for all subjects except those with the most negative Thurstone attitude estimates. This extreme segment of individuals presumably disagreed with the moderately negative items ("Abortion should be illegal except in cases involving incest or rape", "Abortion is basically immoral except when the women's physical health is in danger") because they did not match their extreme positions well enough. The statements were, in essence, too moderate for them. Consequently, these individuals obtained relatively low scores on these items. Their item scores were converted to relatively higher scores after reverse scoring negatively worded items, and thus, the Likert scores made individuals from this segment appear as though they possessed more moderate attitudes.

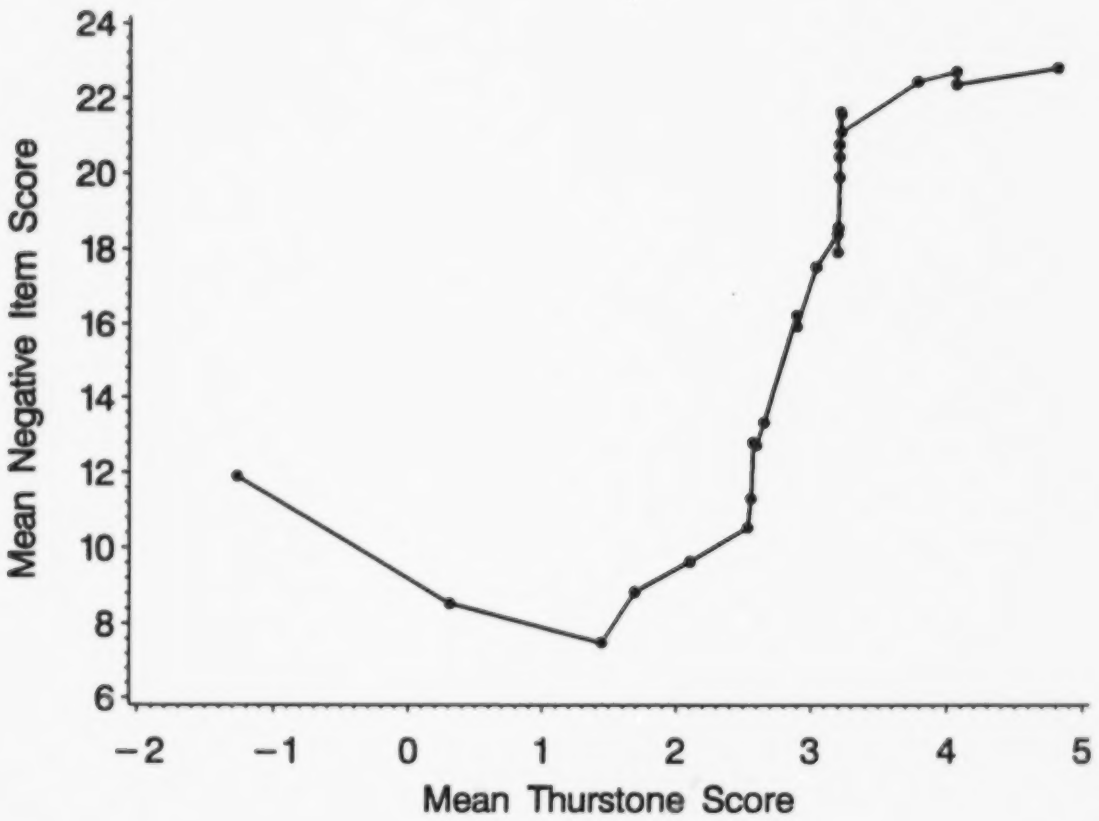

Figure 11. The relationship between mean Likert attitude scores derived from the four negatively worded items and mean Thurstone attitude scores. Means were calculated within relatively homogeneous Thurstone attitude score groups with approximately 30 responses in each group.

The mean Likert scores associated with the positive items are plotted against the mean Thurstone scores from the original test in Figure 12. The 4 positive items had a Cronbach alpha value of .82, and the corrected item-total correlations ranged from .58 to .69. In the case of positively worded items, the Likert and Thurstone scores were more or less monotonically related for all individuals except those with the most extremely positive Thurstone attitude estimates. This segment of individuals agreed less with the moderately positive items ("Abortion should be a woman's choice, but should never be used simply due to its convenience", "Abortion should generally be legal, but should never be used as a conventional method of birth control") presumably because these items were too moderate for them to whole-heartedly endorse. Consequently, these individuals received lower scores on these items, which led to lower Likert scores.

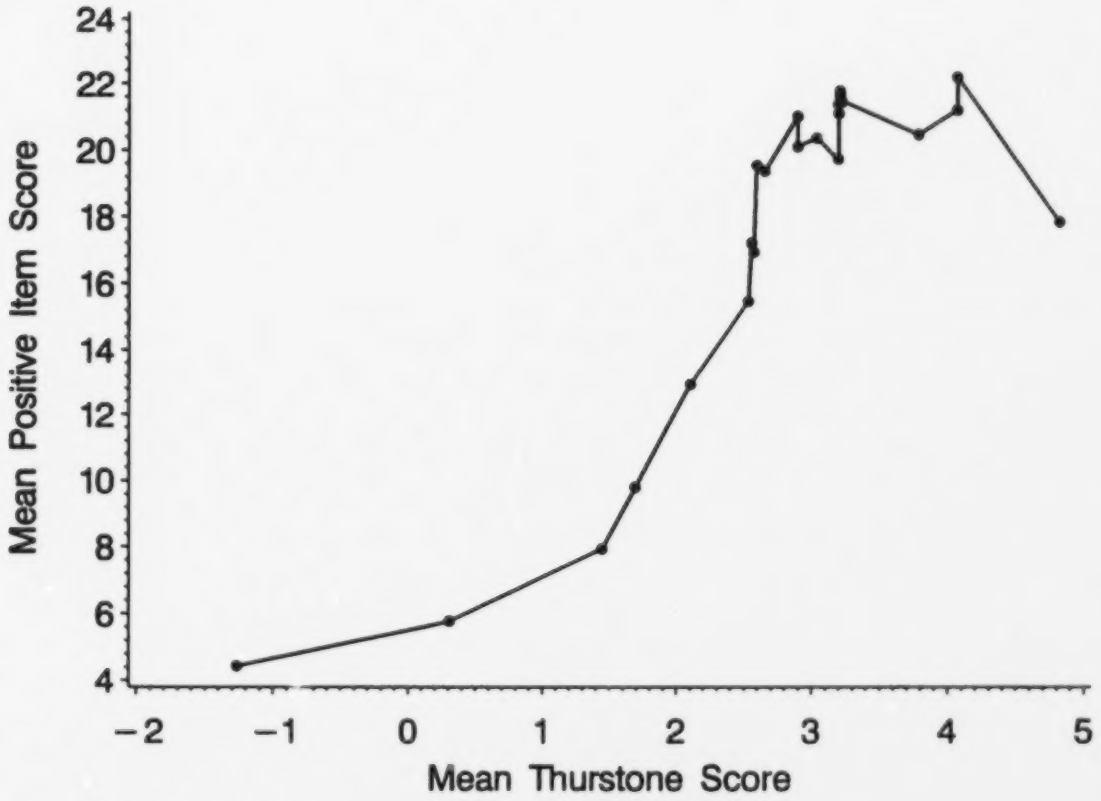

Figure 12. The relationship between mean Likert attitude scores derived from the four positively worded items and mean Thurstone attitude scores. Means were calculated within relatively homogeneous Thurstone attitude score groups with approximately 30 responses in each group.

Discussion and Practical Consequences

The results of the simulation and the real data examples suggest that the Likert procedure may falter for individuals who hold extreme attitudinal positions when responses result from some type of ideal point process. This is because the Likert procedure is functionally a cumulative model of the response process, and as such, it is not always compatible with responses from an ideal point process. In contrast, the Thurstone procedure is functionally an unfolding model, and thus, it does correspond to the situation in which responses follow from an ideal point process. Due to this correspondence, the Thurstone procedure does not suffer from the degraded validity exhibited with the Likert method when individuals with extreme attitudes are measured.

So what is an applied attitude researcher to do? A "knee-jerk" reaction to these data is to include only the most extreme items on a Likert attitude scale - items that are located in the

extreme portions of the attitude continuum and exhibit essentially monotonic characteristic curves. Although this strategy may often work, it suffers from at least two practical drawbacks. First, as shown in the simulation, it may sometimes be difficult to identify the most extreme items with standard item analysis techniques. In these cases, the moderately extreme items with slightly nonmonotonic characteristic curves may appear to be more optimal candidates for the scale as compared to the most extreme items. The degree to which this difficulty is encountered will ultimately depend on a variety of factors which include the distribution of item locations in relation to the locations of individuals and the discriminability of the items. Second, even if one could identify the most extreme items, it may not always be wise to limit the scale solely to them. If items are too extreme (e.g., "Abortion should be a socially acceptable method of birth control", "Abortionists should be harassed"), then few individuals other than those with the most extreme attitudes would endorse them to any appreciable degree. Consequently, the resulting scale scores could vary too little across the sample and fail to adequately differentiate individuals across much of the attitude continuum.

Another overly simplistic response to these results is to simply ignore them! One might rationalize that the Likert procedure generally performed well in a majority of cases from both examples reported in this paper, and it only produced conflicting estimates for a small minority of individuals/simulees studied. However, this rationale is also quite problematic. Individuals with the most extreme attitudes may form a particularly important segment in a given attitude research project. For example, a researcher may want to identify those individuals who like or dislike a given attitude object the most so that such individuals can be compared and contrasted on a variety of potential explanatory variables. Obviously, identification of these individuals could be difficult if the Likert procedure was used to obtain attitude scores. Furthermore, the proportion of persons who are mismeasured with the Likert method will vary from one situation to another and may be more or less than that exhibited in the current examples.

The above results do suggest at least two other courses of action for applied researchers. At the very least, the results suggest that more attention should be devoted to the ICCs associated with scale items. This can only be done if scale values are available for each item, which is usually not the case with traditional Likert scales. Assuming that scale values are developed, then ICCs corresponding to a given sample can be constructed, and the scale can be limited to those items that exhibit essentially monotonic behavior. However, examination of the ICCs should not be confined solely to the scale construction process, but instead, should be part of the scale application regimen. The degree of monotonicity inherent in the ICCs will be highly dependent on the relative locations of persons and items on the attitude continuum, and the range of person locations can obviously change from sample to sample. Therefore, ICCs should be examined in all applied situations where the sample in question is large enough to justify the results.

At the very most, these results suggest that attitude researchers should utilize some type of unfolding model when developing attitude estimates from disagree-agree data. The Thurstone procedure is one example of an unfolding model, but there are many others. For example, a new class of item response models has recently been developed to unfold responses from a disagree-

agree scale. Table 1 classifies several of these models according to whether they assume a particular parametric form for the item response process (i.e., parametric versus nonparametric models) and whether they operate with binary or graded disagree-agree responses. Although some of the models listed in Table 1 require large amounts of data (i.e., samples of 750 subjects or more), not all of them do. For example, when responses fit the Graded Unfolding Model, then as few as 100 subjects responding to a set of 15-20 graded response items can be used to develop accurate estimates of model parameters (Roberts, 1995; Roberts & Laughlin, 1996ab). Moreover, if item scale values are published, then attitude estimates can be obtained from these models on an individual basis. Thus, the models may be quite useful for applied attitude researchers.

Table 1. Unfolding Item Response Theory Models for Disagree-Agree Responses to Attitude Statements.

Model Type	Response Type	Description
Parametric	Binary	Unfolding Threshold Model (DeSarbo & Hoffman, 1986, 1987)
Parametric	Binary	Squared Simple Logistic Model (Andrich, 1988)
Parametric	Binary	PARELLA Model (Hoijtink, 1990, 1991)
Parametric	Binary	Hyperbolic Cosine Model (Andrich & Lou, 1993)
Parametric	Binary or Graded	Graded Unfolding Model (Roberts, 1995; Roberts & Laughlin, 1996ab)
Parametric	Binary or Graded	Generalized Graded Unfolding Model (Roberts, Donoghue & Laughlin, 1996)
Parametric	Binary or Graded	General Hyperbolic Cosine Model (Andrich, 1996)
Nonparametric	Binary or Graded	MUDFOLD Model (van Schuur, 1984, 1993)
Nonparametric	Binary or Graded	Ordinal Scaling Method (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988)

This paper has illustrated a particular problem with the Likert method, but we must emphasize that the problem has only been demonstrated in situations where responses presumably follow from some type of ideal point process. Although disagree-agree responses generally appear consistent with an ideal point process, other response scales need not be. For example, there is no evidence to suggest that frequency responses (e.g., "How often have you picketed a Planned-Parenthood Clinic? - 1=Never, 2=Once, 3=Two to Five Times, 4=More than Five Times") would follow from an ideal point process. In fact, such responses seem intuitively consistent with a dominance process, and thus, should be compatible with cumulative models in general and the Likert method in particular.

References

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*, 33-51.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies, *British Journal of Mathematical and Statistical Psychology, 49*, 347-365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253-276.

Barclay, J. E., & Weaver, H. B. (1962). Comparative reliabilities and the ease of construction of Thurstone and Likert attitude scales. *Journal of Social Psychology, 58*, 109-120.

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12*, 83-97.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

DeSarbo, W. S., & Hoffman, D. L. (1986). Simple and weighted unfolding threshold models for the spatial representation of binary choice data. *Applied Psychological Measurement, 10*, 247-264.

DeSarbo, W. S., & Hoffman, D. L. (1987). Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research, 24*, 40-54.

Edwards, A. L., & Kenney, K. C. (1949). A comparison of the Thurstone and Likert techniques of attitude scale construction. *Journal of Applied Psychology, 30*, 72-83.

Ferguson, L. W. (1941). A study of the Likert technique of attitude scale construction. *Journal of Social Psychology, 13*, 51-57.

Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of Likert scales via confirmatory factor analysis. *Multivariate Behavioral Research, 18*, 275-308.

Green, B. F. (1954). Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology* (1st ed., Vol. 1, pp. 335-369). Cambridge, MA: Addison-Wesley.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55*, 641-656.

Hoijtink (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15*, 153-169.

Jaccard, J., Weber, J., & Lundmark, J. (1975). A multitrait-multimethod analysis of four attitude assessment procedures. *Journal of Experimental Social Psychology, 11*, 149-154.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, No. 140, 5-53.

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology, 5*, 228-238.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Dubuque, IA: Wm. C. Brown.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerishe Mathematik, 10*, 177-183.

Rhoads, R. F., & Landy, F. J. (1973). Measurement of attitudes of industrial work groups toward psychology and testing. *Journal of Applied Psychology, 58*, 197-201.

Roberts, J. S. (1995). *Item Response Theory Approaches to Attitude Measurement*. (Doctoral dissertation, University of South Carolina, Columbia, 1995). *Dissertation Abstracts International, 56*, 7089B.

Roberts, J. S., & Laughlin, J. E. (1996a). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.

Roberts, J. S., & Laughlin, J. E. (1996b). *The graded unfolding model: A unidimensional item response model for unfolding graded responses* (RR-96-16). Princeton, NJ: Educational Testing Service.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1996). A generalized item response model for unfolding responses from a graded scale. Paper presented at the 61st annual meeting of the Psychometric Society, June 28, Banff, Alberta, Canada.

Safir, M. A. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika, 2*, 179-198.

Seiler, L. H., & Hough, R. L. (1970). Empirical comparisons of the Thurstone and Likert techniques. In G. Summers (Ed.), *Attitude measurement*, (pp. 159-173). Chicago: Rand McNally.

Thurstone, L. L. (1927a). Psychophysical analysis. *American Journal of Psychology, 38*, 368-389.

Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Thurstone, L. L. (1927c). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 26*, 249-269.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church*. Chicago: University of Chicago Press.

van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists*. Amsterdam: CT Press.

van Schuur, W. H. (1993). Nonparametric unidimensional unfolding for multicategory data. *Political Analysis, 4*, 41-74.

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model can be used instead. *Applied Psychological Measurement, 18*, 97-110.